Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions

LUCA DELLE MONACHE

National Center for Atmospheric Research, Boulder, Colorado

THOMAS NIPEN

University of British Columbia, Vancouver, British Columbia, Canada

YUBAO LIU AND GREGORY ROUX

National Center for Atmospheric Research, Boulder, Colorado

ROLAND STULL

University of British Columbia, Vancouver, British Columbia, Canada

(Manuscript received 6 October 2010, in final form 16 February 2011)

ABSTRACT

Two new postprocessing methods are proposed to reduce numerical weather prediction's systematic and random errors. The first method consists of running a postprocessing algorithm inspired by the Kalman filter (KF) through an ordered set of analog forecasts rather than a sequence of forecasts in time (ANKF). The analog of a forecast for a given location and time is defined as a past prediction that matches selected features of the current forecast. The second method is the weighted average of the observations that verified when the 10 best analogs were valid (AN). ANKF and AN are tested for 10-m wind speed predictions from the Weather Research and Forecasting (WRF) model, with observations from 400 surface stations over the western United States for a 6-month period. Both AN and ANKF predict drastic changes in forecast error (e.g., associated with rapid weather regime changes), a feature lacking in KF and a 7-day running-mean correction (7-Day). The AN almost eliminates the bias of the raw prediction (Raw), while ANKF drastically reduces it with values slightly worse than KF. Both analog-based methods are also able to reduce random errors, therefore improving the predictive skill of Raw. The AN is consistently the best, with average improvements of 10%, 20%, 25%, and 35% with respect to ANKF, KF, 7-Day, and Raw, as measured by centered root-mean-square error, and of 5%, 20%, 25%, and 40%, as measured by rank correlation. Moreover, being a prediction based solely on observations, AN results in an efficient downscaling procedure that eliminates representativeness discrepancies between observations and predictions.

1. Introduction

In recent years, the increasing demand for accurate weather forecasts has led to a steady improvement of the skill of numerical weather predictions at both global and regional scales. Despite these improvements, such predictions are still affected by imperfect initial conditions, numerical approximations, and simplification (or altogether lack of representation) of the physical and chemical

DOI: 10.1175/2011MWR3653.1

processes that govern the evolution of the atmosphere. These imperfections, approximations, and simplifications result in random and systematic errors (e.g., bias) that affect the predictions' accuracy. Bias here is defined as the "difference of the central location of the forecasts and the observations" (Jolliffe and Stephenson 2003).

Several contributions can be found in the literature proposing algorithms to predict these errors, particularly the systematic component: methods based on 1) runningmean corrections for numerical weather predictions (NWP; e.g., Stensrud and Skindlov 1996; Stensrud and Yussouf 2003; Eckel and Mass 2005; Hacker and Rife 2007), and air quality (AQ) predictions (e.g., Wilczak et al. 2006); 2) state-dependent corrections (e.g., Leith 1978;

Corresponding author address: Luca Delle Monache, Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000. E-mail: lucadm@ucar.edu

DelSole and Hou 1999; Danforth et al. 2007; Danforth and Kalnay 2008); 3) postprocessing algorithms inspired by the Kalman filter (KF) for NWP (e.g., Homleid 1995; Roeger et al. 2003; McCollor and Stull 2008; Rincon et al. 2010) and AQ (Delle Monache et al. 2006, 2008; Djalalova et al. 2010; Kang et al. 2010); 4) model output statistics approaches (e.g., Glahn and Lowry 1972; Carter et al. 1989; Jacks et al. 1990; Hart et al. 2004; Wilks and Hamill 2007); and 5) gene-expression algorithms (e.g., Bakhshaii and Stull 2009).

The KF postprocessing approach implemented in this study is linear, adaptive, and recursive; it is easy to implement and computationally inexpensive. Its advantages include a short training period (i.e., few weeks), and the ability to adapt to changing synoptic conditions, changing seasons, and even changing models. However, a disadvantage of this method is that it is not likely to predict sudden changes of the forecast error caused by rapid transitions from one weather regime to another. Initially KF may not be able to correctly estimate the forecast error when these changes happen, but it then quickly adapts to these changes in the following cycles (Delle Monache et al. 2006).

Two new methods are proposed to address this shortcoming. The first method combines the KF postprocessing algorithm with an analog approach (ANKF). The second scheme is based purely on observations that verified when analog forecasts were valid (AN). The analog of a forecast for a given location and time is defined as a past prediction that matches selected features of the current forecast. The basic idea underlying the new proposed methods is that if forecasts in the past (called here *analogs*) can be found that are similar to the current prediction, we can infer information about the forecast error by analyzing the errors of the analogs, for which verifying observations are available.

The analog concept and its potential usefulness for weather forecasting have been explored by several investigators. Various procedures have been formulated, including different predictors and analog-selection criteria. Applications include idealized cases with low-order models (Ren and Chou 2006; Ren et al. 2009), general circulation modeling (Radinovic 1975; Van den Dool 1989, 1994, 2007; Gao et al. 2006; Ren and Chou 2007); longrange weather (Bergen and Harnack 1982; Livezey and Barnston 1988; Toth 1989; Xavier and Goswami 2007), sea ice-anomaly prediction (Chapman and Walsh 1991), shortterm visibility (Esterle 1992), mesoscale transport forecasts (Carter and Keislar 2000), El Niño–Southern Oscillation index forecasts (Drosdowsky 1994), and calibration of probabilistic predictions (Hamill and Whitaker 2006).

ANKF and AN are tested with 10-m wind speed forecasts from the Weather Research and Forecasting (WRF) modeling system (Skamarock et al. 2008), with 1–24-h forecasts issued at 0100 UTC. Hourly predictions are compared to 400 surface wind observations for a 6-month period and over a domain centered on northeast Colorado. The performance of ANKF and AN methods is compared to the skill of 1) the raw forecast (Raw), 2) a 7-day running-mean correction (7-Day), and 3) the KF approach.

The following section describes the postprocessing methods, section 3 outlines the design of the testing procedure, section 4 includes a sensitivity analysis for ANKF and AN, and section 5 presents the results. Section 6 summarizes this work and states conclusions drawn from it.

2. Postprocessing methods

First, a simple 7-day running-mean correction (7-Day), used as a reference against which the more sophisticated methods are compared, is introduced. Then, the KF method is described with a brief discussion of its advantages and disadvantages. Two new methods are introduced to address the disadvantages of the KF method: ANKF that consists in running KF through an ordered set of analog forecasts rather than in time and AN, which is solely based on analogs as explained in section 2d. This section is concluded with an example of AN and ANKF, and a summary of key aspects of their design and differences.

a. 7-day running-mean correction (7-Day)

The most recent 7-day running-mean of the prediction errors is used as a bias estimate for the current forecast (e.g., Wilczak et al. 2006 for AQ; and Stensrud and Skindlov 1996; Stensrud and Yussouf 2003; Eckel and Mass 2005; Hacker and Rife 2007 for NWP). We choose a 7-day averaging window after investigating the sensitivity of the performance of the running-mean with the dataset analyzed in this study (not shown), as measured by root-mean-square-error (RMSE) and rank correlation (defined in section 3b). Similar window length has also been chosen by other authors for different applications (e.g., Stensrud and Skindlov 1996; Stensrud and Yussouf 2003; Wilczak et al. 2006; Hacker and Rife 2007). The advantage of this method is its simplicity and ability to improve raw predictions, while its limitations stem from its lack of skill in predicting bias changes occurring at a temporal scale shorter than 7 days.

b. Postprocessing method inspired by the Kalman filter

In the postprocessing method inspired by the KF, past prediction errors at a given location (i.e., the difference between forecasts and observations) are used to estimate the bias in the current raw forecast [see Fig. 1 in Delle Monache et al. (2006) for the algorithm flow diagram]. As opposed to the 7-day running mean, it can be seen as an exponential weighted mean where recent errors are weighted more than past errors.

The true (unknown) forecast bias x_t at time *t* is modeled by the previous true bias plus a noise term η_t , which is normally distributed with zero-mean and a variance $\sigma_{n,t}^2$ (Bozic 1994):

$$x_t = x_{t-\Delta t} + \eta_t, \tag{1}$$

where Δt is a time lag. Because of unresolved terrain features, numerical noise, lack of accuracy in the physical parameterizations, and errors in the observations themselves, the KF approach further assumes that the forecast error y_t (forecast minus observation at time t) is corrupted by an unsystematic error term ε_t :

$$y_t = x_t + \varepsilon_t = x_{t-\Delta t} + \eta_t + \varepsilon_t, \qquad (2)$$

where ε_t is normally distributed with zero-mean and variance $\sigma_{e_t}^2$.

In the postprocessing algorithm inspired by the Kalman filter (Kalman 1960) the recursive predictor of x_t (derived by minimizing the expected mean-square error) can be written as a combination of the previous predicted bias and the previous forecast error:

$$\hat{x}_{t+\Delta t|t} = \hat{x}_{t|t-\Delta t} + K_t(y_t - \hat{x}_{t|t-\Delta t}),$$
 (3)

where the hat (^) indicates the estimate.

The weighting factor K_t , called the Kalman gain, can be calculated from

$$K_t = \frac{p_{t-\Delta t} + \sigma_{\eta,t}^2}{\left(p_{t-\Delta t} + \sigma_{\eta,t}^2 + \sigma_{\varepsilon,t}^2\right)},\tag{4}$$

where *p* is the expected mean-square error, which can be computed as follows:

$$p_t = (p_{t-\Delta t} + \sigma_{\eta,t}^2)(1 - K_t).$$
 (5)

The KF algorithm will quickly converge for any reasonable initial estimate of p_0 and K_0 . It is easy to implement and fast running, requiring storage of few coefficients for each site and forecast hour. The filter algorithm is run independently on data for each forecast hour, using only values from previous days at the same forecast hour [i.e., $\Delta t = 24$ h in Eqs. (1)–(5)] to take into account the diurnal behavior of the forecast bias. Additional details of the algorithm including how $\sigma_{n,t}^2$ and $\sigma_{e,t}^2$ are estimated can

be found in section 2 and appendix A of Delle Monache et al. (2006).

The KF approach adapts its coefficients as new forecasts and observations become available. Advantages are a short training period (i.e., few weeks), and the ability to adapt to changing synoptic conditions, changing seasons, and even changing weather forecast or AQ models. A disadvantage is that it is less likely to predict extreme bias events; namely, it is unable to anticipate a large forecast error when all errors for the past few days have been smaller.

c. Kalman filtering through an ordered set of analog forecasts (ANKF)

The disadvantages of KF not being able to predict large day-to-day changes in the prediction error have prompted the development of a new method, where KF is run through an ordered set of analog forecasts rather than in time. Analogs here are defined as past forecasts that are similar to the current forecast (as measured by a metric defined later) for which a correction in real time is desired. Figure 1a shows the KF deficiency of producing an accurate estimate of the current forecast error when the error behavior constantly changes. The time series shown are synthetic data generated for illustration purposes. Observations (open circles) show a daily sequence of alternating wind regimes, where calm situations are followed by windier days. The raw model prediction in this example tends to have larger biases when winds are calmer. The forecast is issued today during high wind conditions, and the prediction is for lower winds. Being a recursive procedure, KF uses all the available information to estimate the error of the current forecast, giving higher weights to the most recent data. In the cartoon example for the last forecast, KF weights more days when errors were low given the windy conditions, resulting in a correction that underestimates the current forecast error (as indicated by the white arrow).

A way to account for this problem is to run KF through an ordered set of analog forecasts rather than in time, as illustrated in Fig. 1b. The analogs to the current forecast are ranked (left to right) from the farthest (worst analog) to the closest (best analog). By running KF through the ordered (from worst to best) analogs the correction for the current forecast gives more weight to the analog forecasts closer to it. This results in a better correction as can be seen by comparing the white arrows at the forecast time (far right) in Figs. 1a,b.

If proper analogs are defined and found, such a procedure can cope with drastic changes in the forecast error. For this reason, the key aspect of such an approach is to determine a suitable metric that ranks past forecasts by how similar they are to the current forecast.



FIG. 1. Schematic representation of a Kalman filter correction for wind speed prediction (WSPD) (a) run in time (KF) or (b) through an ordered set of analog forecasts (ANKF). White arrows at forecast time (far right) indicate the postprocessing methods estimate of the forecast error. Circles indicate observations, asterisks refer to the raw prediction, and the dashed line represent the corrected predictions.

That is, past predictions that are very similar to the current forecast should also exhibit very similar errors. Such a metric is defined as follows:

$$\|F_{t}, A_{t'}\| = \sum_{i=1}^{N_{v}} \frac{w_{i}}{\sigma_{f_{i}}} \sqrt{\sum_{j=-\tilde{t}}^{\tilde{t}} (F_{i,t+j} - A_{i,t'+j})^{2}}, \qquad (6)$$

where F_t is the forecast to be corrected at the given time tand station location; $A_{t'}$ is an analog forecast at a time t'before F_t was issued and at the same location, N_v and w_i are the number of physical variables and their weights, respectively; σ_{f_i} is the standard deviation of the time series of past forecasts of a given variable at the same location; \tilde{t} is an integer equal to half the width of the time window over which the metric is computed; and $F_{i,t+j}$ and $A_{i,t'+j}$ are the values of the forecast and the analog in the time window for a given variable.

Analogs are searched across multiple physical variables and over a time window for a given location and forecast time [Eq. (6)]. The idea is to find past forecasts that were predicting similar values and temporal trends for the forecasted quantity (wind speed in this study), and for variables that exhibit correlations to the quantity of interest (e.g., wind speed itself, wind direction, pressure, etc.). The assumption is that if these forecasts are found, their errors will likely be similar to the error of the current forecast, which can be inferred from them. If some quantities are known to be more correlated to the quantity to be corrected, then they can be assigned a higher weight (w_i). The terms in the summation in Eq. (6) are normalized with σ_{f_i} to allow for the combination of physical variables with different units as well to avoid that the metric value is dominated by a single variable. The weight w_i assigned to each of the variables used in Eq. (6) to search for analogs is set to 1. No attempt was made to find optimal values for w_i .

d. Weighted analogs (AN)

The AN forecast is the weighted average of the observations that occurred when analog forecasts were valid:

$$AN_{t} = \sum_{i=1}^{N_{a}} \gamma_{i} OA_{i,t_{i}}, \qquad (7)$$

where AN_t is the AN forecast at time t at a given location, N_a is the number of analogs, $\{OA_{i,t_i}\}_{i=1,2,...,N_a}$ are the verifying observations of the best N_a analogs as measured by the metric defined in Eq. (6), and t_i are times when these analog forecasts were issued (earlier than when the



FIG. 2. Example of a set of analogs selected for a 10-m wind speed forecast issued at a surface station in central Montana (at 44.69°N, 111.10°W) at 0100 UTC 25 Sep 2009. Observation of (a) wind speed (WSPD, m s⁻¹), (b) prediction of WSPD, (c) wind direction (WDIR, °), (d) surface pressure (*P*, hPa), (e) humidity (*Q*, g kg⁻¹) and (f) 10-m temperature (*T*, K). The solid black circle and horizontal black line indicate the variable value at forecast time (vertical dashed line). The 10 best analogs are also shown, indicated by the shaded circles along with each analog rank. The darker the shading, the better the analog.

forecast to be corrected was issued). The weight associated with each analog (γ_i) is computed as follows:

$$\gamma_{i} = \frac{\frac{1}{\|(F_{i}, A_{i, t_{i}})\|}}{\sum_{j=1}^{N_{a}} \frac{1}{\|(F_{t}, A_{j, t_{i}})\|}},$$
(8)

and is proportional to the inverse of the distance of the analog from the forecast [Eq. (6)] and normalized with the sum of the inverse of this distance computed for each analog; the shorter the distance between an analog and a forecast, the higher the weight that will be assigned to the observation that verified when the analog was issued. The weights sum to 1.

e. Example of AN and ANKF

Figure 2 shows an example of a set of analogs selected for a 10-m wind speed forecast issued at a surface station in central Montana at 0100 UTC 25 September 2009. Figure 2a shows a 4-month time series of observed wind speed (WSPD), with the verifying values (not available at forecast time) shown with the black circle at the far right of the plot and a continuous horizontal line. Figure 2b shows the time series of the raw model prediction of WSPD, with the current forecast (to be corrected) indicated by the solid black circle at the far right. For this example, analogs were found by exploring the time series of wind speed (Fig. 2b) and direction (WDIR, Fig. 2c), surface pressure (P, Fig. 2d), humidity (Q, Fig. 2e), and 10-m temperature (T, Fig. 2f). Similar to Figs. 2a,b, in Figs. 2c-f the value at the forecast time (vertical dashed line in each panel) is indicated by the solid black circle and continuous horizontal line.

Also shown in this example are the 10 best analogs, indicated by the shaded circles, along with each analog rank. The shading indicates the quality of the analog: the darker the better. The three best analogs are close in time to the current forecast; however, the most recent forecast is not included in the 10 best analogs. Moreover, the analogs from the fourth to the ninth ranking are found in early May (i.e., four months before the current forecast), while the tenth analog is more than a month older than that. For the forecast to be corrected at this location, issued at the end of September, the 10 best analogs are found either a few days before, or toward the end of spring, when the meteorological conditions were closer to those of the current forecast than the conditions found during the months of June, July, or August, that were hotter and more humid (as shown in Figs. 2e,f).

The observed wind speed value at the forecast time is 3.1 m s^{-1} (Fig. 2a, solid black circle at the far right), while the raw forecast value is 5.2 m s^{-1} (Fig. 2b, solid black circle at the far right), which is corrected by AN, and ANKF to 3.6, and 4.6, respectively (not shown). For its estimate AN does not use three out of six of the most recent forecasts, because they do not closely match the forecast to be corrected, while ANKF uses them but with lower weights. The differences among the two analog-based methods described in the following section may explain the difference in performance.

f. Key aspects of AN and ANKF

Important aspects of the new analog-based postprocessing procedures can be summarized as follows:

- Analogs are searched in forecast space only (i.e., no observations are used to select the best analogs), and across different meteorological variables that exhibit correlations to the quantity of interest.
- · Analogs are searched independently for each forecast time and location; the assumption is that when the space and period of the day from which analogs are selected is reduced (in other words, by "localizing" the search procedure both in space and time), also the degrees of freedom of finding matching forecasts are reduced, resulting in a more tractable problem. This can be contrasted with approaches where full three-(space) or four- (space and time) dimensional fields are compared while searching for analogs. Other authors have presented analog-based approaches including a localization strategy in space and time (Van den Dool 1989, 1994; Hamill and Whitaker 2006). Van den Dool (1989) reported that while previous attempts to find analogs by matching large scale flow patterns were discarded in the literature, an important factor contributing to the success of his localized analogbased method was "the lowering of the degrees of freedom in finding matching states."
- ANKF and AN are both based on a weighted contribution of the analogs, where the weights are proportional

to the quality of the analogs. However, while ANKF assigning these weights by running KF [i.e., with the Kalman gain as defined in Eq. (4)] through the ordered (from worst to best) analogs, AN directly estimates each analog weight as directly proportional to his closeness to the forecast [Eq. (8)].

- While KF and therefore ANKF is not optimal when the assumptions of Gaussianity and linearity (section 2b) are not met (as likely the case with the dataset presented here including predicted and observed wind speed), AN does not have these assumptions built in.
- While ANKF tries to estimate the current forecast error, that is then subtracted from it, AN is based purely on a linear combination of the observations that verified when the best analog forecasts were valid.

3. Experiment description

a. NWP predictions

The postprocessing methods described in section 2 are applied to 10-m wind speed forecasts from an operational mesoscale forecasting system based on the WRF model (Skamarock et al. 2008). While the system provides 8 forecast cycles per day, the only forecasts applied in this study are the 1-24-h forecasts (at 1-h increments) issued at 0100 UTC. The operational WRF system is run over the western United States with three nested domains centered over northeast Colorado having 30-, 10-, and 3.3-km horizontal grid increments and 37 vertical levels (12 levels are located in the lowest 1 km). The parameterizations chosen for these experiments include the Purdue Lin microphysics scheme, the Yonsei University planetary boundary layer (PBL) scheme, Monin-Obukhov for the surface layer, Kain-Fritsch for the convective processes (only in the two coarser domains), and the Noah land surface model for the land surface scheme.

The model output postprocessing methods described above are tested with observations from 400 surface stations for a 6-month period (from 3 May 2009 to 8 November 2009). The stations are located within the finest inner domain, and their spatial distribution covers different topographical and land-use type: the central Rocky Mountain with high elevation and complex topography; the plains to the east of it; a few dry desert areas to the south; and urban, suburban, and rural locations. A 1-yr-long independent dataset (described in section 5d) is also used to study the effects of using datasets with different lengths on the analog-based methods' performance.

b. Metrics

The following metrics are computed to evaluate the performance of the different postprocessing methods.

Following Taylor (2001), RMSE can be decomposed as follows:

$$RMSE^{2} = \frac{1}{N_{p}} \sum_{i=1}^{N_{p}} (F_{i} - O_{i})^{2} = CRMSE^{2} + BIAS^{2}, \quad (9)$$

where

$$CRMSE = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} \left[(F_i - \overline{F}) - (O_i - \overline{O}) \right]^2}, \quad (10)$$

and

$$BIAS = \overline{F} - \overline{O}.$$
 (11)

The quantity N_p is the number of available (F_i, O_i) pairs with F_i and O_i being a forecast and observation at the same time and location, and \overline{F} and \overline{O} are the prediction and observation averages over the N_p values. The centered root-mean-square error (CRMSE) is considered to be the random component of RMSE, while the systematic component is represented by BIAS. The terms CRMSE² in Eq. (9) is equal to $s_f^2 + s_o^2 - 2s_f s_o r_{fo}$ (Murphy 1988) where s_f and s_o are the standard deviations of the marginal distributions of the forecast and observations, respectively, and $r_{\rm fo}$ is the Pearson correlation coefficient between forecast and observations (hereafter referred to as "correlation"). We compute also the normalized standard deviation (NSD), as s_f/s_o . Finally, we evaluate the Spearman correlation coefficient that can be thought of as the Pearson correlation coefficient between the ranked variables, and is referred to here as the "rank correlation" (Wilks 2006).

We include the above metrics in this analysis for the following reasons:

- CRMSE gives an indication of errors other than bias, and it can be associated with the intrinsic predictive skill of the forecast that can be limited by the coarse or nonexistent representation of specific physical processes.
- BIAS estimates the systematic errors whose sources may be the model misrepresentation of topography or coastline complexity, offset parameter values, biased initial conditions and inputs in general, etc.
- Correlation and NSD are used here to generate Taylor's diagrams (Taylor 2001). These plots allow us to assess the degree of pattern correspondence between the predictions and the observations.
- Rank correlation is a nonparametric (i.e., distribution free) statistic that measures the strength of the

monotone associations between two variables, allowing for a nonlinear relationship between the predictions and observations. It is appropriate when the quantity of interest, wind speed in this study, exhibits a non-Gaussian distribution. It is a robust and resistant alternative to the Pearson correlation.

4. Sensitivity analysis for the analog-based methods

The sensitivity to a number of parameters and implementation options in Eqs. (6) and (7) for the new methods ANKF and AN is presented. This is followed by an analysis of the effects of using different quantities to search for analogs.

The results of the sensitivity analysis can be summarized as follows:

- Values for *t̃* (the half-width time windows over which squared differences between analog and forecast values are computed for a given location) in the set {*i*}_{*i*=0,1,...,12}, *i* ∈ N were tested, resulting in minimal differences between the different runs, with a value of *t̃* = 1 producing the best results. For this reason, *t̃* = 1 was used. Because the data used here have an hourly frequency, this corresponds to a comparison of the analog and prediction over a 2-h window, that is able to capture the relevant information in terms of the predicted value and its trend, based on the results shown in section 5. Different datasets may have a different optimal value for *t̃*.
- The N_a values from 1 to 15 were tested; the performance of the methods levels off around $N_a = 10$, and then slowly drops off. Therefore, N_a has been set to 10; this number is likely dependent on the dataset length (see section 5d) and the available physical variables to search for analogs, and cannot be generalized for other datasets and different applications of the analog-based methods.
- We let t' in Eq. (6) span any hour of the past days while searching for analogs. The best results were obtained when t' was equal to the hour of the forecast to be corrected, meaning that the diurnal behavior of errors is an important aspect of forecast inaccuracy (as also discussed in section 5b). For this reason we choose t' equal to the forecast hour to be corrected, which also produces a faster running algorithm than when t' is allowed to vary across all the hours of the day.

Next, we examine the effects of using different physical variables to search for analogs. Figure 3 shows the results of this analysis: Figs. 3a,b show CRMSE and rank correlation computed with N_p equal to all the available pairs (in space and time) of observations and predictions



FIG. 3. (a) CRMSE (m s⁻¹), (b) rank correlation, and (c) physical variable used to search for analogs against 17 sensitivity runs. In (a) and (b), solid black bars refer to the results of the post-processing method based on running KF through an ordered set of analog forecasts (ANKF), and solid white bars refer to the results to the method solely based on analogs (AN).

for a set of 17 sensitivity runs; Fig. 3c shows which variables were used in the search for analogs in each run [i.e., the variables used to compute the summation in Eq. (6)]. These include all the available variables in this dataset.

The ANKF (black solid bars) performs best when only WSPD is used to search for analog (run 1), while the best run for AN (white solid bars) is run 10 (WSPD, WDIR, and P). The different behavior of the two methods is expected, given the differences between the approaches as explained in section 2f. Both methods perform well with the run 10 configuration (WSPD, WDIR, and P), and this is the option used for the results showed in section 5. The inclusion of P rather than T or Q is meaningful also from a physical point of view, given the stronger relationship between WSPD and P rather than with T and Q. Other variables that were not available in this dataset, but that could have been useful while searching for analogs for WSPD may have been variables carrying information about the atmospheric stability at the location of interest, such as the Monin-Obukhov length or the PBL depth, or precipitation and cloud cover.

5. Results

The performances of the postprocessing methods are tested with the NWP model runs and the metrics described in section 3. First the metrics are computed with all the available data. Then the same metrics are computed with all the available data at a given time or point in space. This section is concluded with an analysis of the benefits of ANKF and AN with respect to KF when using datasets with a different length in time.



FIG. 4. (a) Improvements (%) of the postprocessing methods relative to the raw forecast (Raw) as a function of the magnitude of the day-to-day variation of forecast absolute error (E_{day} , m s⁻¹). The methods are the 7-day running-mean correction (7-Day, solid black circles), the KF (solid white diamonds), KF run through an ordered set of analog forecasts (ANKF, solid white squares), and the method based only on analogs (AN, solid white circles). (b) Counts over all the stations and the period analyzed in this study of the binned magnitude of the day-to-day variation of E_{day} (increments of 0.25 m s⁻¹).

a. Global metrics

Here the metrics are computed with N_p equal to the number of pairs of observations and predictions available across all stations and times. First, we checked if the analog-based methods are doing what they have been designed for. Figure 4a displays the improvements of 7-Day (black circles), KF (white diamonds), ANKF (white squares), and AN (white circles) methods relative to raw NWP output (Raw), as a function of the magnitude of the day-to-day variation of the raw forecast absolute error (E_{day}) . The same symbols for each method will be used in the following figures (with solid five-point stars associated with Raw) unless otherwise stated. Figure 4b shows the counts of $|E_{day} - E_{day-1}|$ binned in increments of 0.25 m s^{-1} . The relative improvement of KF and 7-Day methods with respect to Raw starts at 40% and 30%, respectively, when there is no day-to-day change of the forecast-error magnitude. But as $|E_{day} - E_{day-1}|$ grows, the improvements rapidly drop and become negative for $|E_{day} - E_{day-1}|$ bigger than 3 m s⁻¹. This confirms the weakness of both methods as discussed in sections 2a,b. Because the 7-Day method is slower to adapt to forecast-error changes, its performance is inferior to KF. For the majority of the cases (as shown by the counts in Fig. 4b) the application of both methods results in an improvement over the raw model output.

VOLUME 139

ANKF also exhibits a decreasing improvement with respect to Raw with increasing $|E_{day} - E_{day-1}|$, but with a much gentler slope than the KF or 7-Day method. ANKF always leads to positive improvements going from about 38% when $|E_{day} - E_{day-1}|$ is equal to 0, down to 15% when $|E_{day} - E_{day-1}|$ is equal to 8 m s⁻¹. In contrast, AN manifests an increase in improvement with respect to Raw with increasing $|E_{day} - E_{day-1}|$, going from 30% with no error change up to 40% when $|E_{dav} - E_{dav-1}|$ is close to its maximum, 8 m s^{-1} .

Both ANKF and KF are better than AN when $|E_{day} E_{\text{dav-1}}$ is lower than roughly 1.0 m s⁻¹. However, when such error grows larger, AN gets quickly better than the KF-based methods, with quite large improvements with respect to both. This means that ANKF tends to weight higher than AN the contribution from the best analog (and KF weights high the most recent forecast). Being that both ANKF and KF are based on a recursive procedure, they can be seen as an exponential weighted mean as explained in section 2b. However, this exponential weighting strategy of the KF-based methods pays off only when the best analog is an excellent match (as when the day-to-day error change little); but AN has an advantage in all the other cases, being its weights directly proportional to their closeness to the forecast, and therefore able to account for their quality more effectively. The counts in Fig. 4b show that ANKF is better than AN roughly 40% of the times, and this contributes to the explanation of the overall performance differences between the two methods presented in the reminder of this section.

The results shown in Fig. 4 underline the benefits of the analog concept: if a correction is done by inferring the current forecast error from past forecasts with similar characteristics, the quality of such correction does not suffer from sudden changes in model error, as long as similar forecasts were recorded in the past.

A concise way to display and study the metrics is to use a Taylor diagram. It can be used to create a multistatistic plot of correlation, CRMSE, and NSD (Fig. 5), and it allows us to estimate the degree of pattern correspondence between the predictions and the observations (Taylor 2001). Correlation and NSD are computed for each method and the observations, and are then plotted on the diagram with polar coordinates defined by the pair (correlation and NSD). The CRMSE of a given method corresponds to the distance on the diagram between the marker representing the method and the one representing the observations (black square). The AN is the closest to the observations, followed by ANKF, KF, 7-Day, and the Raw prediction. Raw has the best NSD, close to the perfect value of 1 (i.e., when the standard deviation of the forecasts is equal to the standard deviation of the observations), while 7-Day and KF



FIG. 5. Taylor diagram showing the raw forecast (Raw, 5-point solid white star), and the postprocessing methods: the 7-day runningmean correction (7-Day, black circles), the KF (white diamonds), KF run through an ordered set of analog forecasts (ANKF, white squares), and the method based only on analogs (AN, white circles). The azimuthal position gives the correlation (straight gray lines), while the radial distance from the origin is proportional to the normalized standard deviation (circular gray lines). The black square represents the observations. The distance between the observation and a given point (black circular lines) is proportional to the CRMSE between the observations and the forecast having the correlation and standard deviation of the given point.

overestimate the standard deviation of the observations, and ANKF and AN underestimate it. However, all postprocessing methods improve the correlation with observations when compared to Raw, resulting in points closer (i.e., with lower CRMSE) to the solid black square corresponding to the observation.

Given the relationship expressed by Eq. (9), if CRMSE is plotted against bias, then the distance of a point on that diagram from the origin is equal to RMSE. Figure 6 illustrates such a diagram for the raw forecast and the postprocessing methods, where the gray circular lines have a radial distance from the origin equal to the RMSE of each prediction. All methods perform well in terms of reducing the bias of Raw, AN being the best, followed by KF, ANKF and 7-Day. All methods also reduce CRMSE, with AN being again the best, followed by ANKF, KF, 7-Day, and then Raw. The latter is also the ranking of the methods when RMSE is the metric considered.

It is important to notice that these methods not only reduce the bias component of RMSE, but also the remaining portion of it. The former can be seen as the ability of the methods to reduce the difference of the



FIG. 6. Bias (m s⁻¹) as a function of CRMSE (m s⁻¹), showing the raw forecast (Raw, 5-point white star), and the postprocessing methods: the 7-day running-mean correction (7-Day, black circles), the KF (white diamonds), KF run through an ordered set of analog forecasts (ANKF, white squares), and the method based only on analogs (AN, white circles). Gray circular lines have a radial distance from the origin equal to the RMSE values of each prediction.

central location of the forecasts and the observations, while the latter results from the ability to add predictive skill to the raw forecast by reducing random errors. The AN is based purely on the verifying observations of the analogs, that provide physically based insight on the atmospheric state, improving the prediction skill. Moreover, using past observations to produce the prediction as with AN, results in an efficient downscaling procedure that eliminates representativeness issues associated with the fact that the observations are typically instantaneous (or averaged over a brief period of time) at a given location, while models provide averages over volumes and time periods corresponding to the computational domain grid boxes and integration time step, respectively.

b. Metrics evaluation as a function of time

In this section the metrics are computed with all the available data at a given time (i.e., N_p equal to the number of pairs of observations and predictions available at a given time and across all the stations), and for each of the 24 forecast hours available. Figures 7, 8, and 9 show the temporal variation of CRMSE, bias, and rank correlation, respectively. The shaded areas (light gray) in these figures correspond to nighttime hours, where sunrise–sunset times are estimated for the domain central location as an average of the daily values over the experiment period. The 95% bootstrap confidence interval of the computed statistics is shown by the shaded areas in darker gray underneath the markers representing the different methods.

The collapse of the PBL is often a challenging process to be predicted, and that is reflected by the jump in CRMSE values for all methods right after sunset, even though this increase in random errors is much less pronounced for the analog-based methods when compared to Raw, 7-Day, and KF (Fig. 7). This is another indication



FIG. 7. Temporal evolution of CRMSE (m s⁻¹) for the raw forecast and postprocessing methods across the 24 h of forecast (issued at 0100 UTC): raw forecast (Raw, 5-point white star), the 7-day running-mean correction (7-Day, black circles), the KF (white diamonds), KF run through an ordered set of analog forecasts (ANKF, white squares), and the method based only on analogs (AN, white circles). Nighttime hours are shaded in light gray (sunrise–sunset times are estimated for the domain central location as an average over the experiment period). The 95% bootstrap confidence interval of the computed statistics is shaded in dark gray underneath the method markers.

of the analog-based methods' ability to improve the predictive skill. Throughout the night, the CRMSE values stay constant or even decrease. Around sunrise, all methods see an increase in CRMSE, given the longer forecast time and the uncertainly associated with the prediction of the PBL growth during daytime. Based on this metric, ANKF and AN show improvements across all the forecast hours with respect to KF in the range of 10% and 20%, respectively; and with even larger



FIG. 8. As in Fig. 7, but for bias (m s^{-1}).



FIG. 9. As in Fig. 7, but for rank correlation.

improvements with respect to Raw and 7-Day with values around 25% and 35%, respectively. The AN is consistently the best across the forecast period, showing statistically significant improvements over ANKF on the order of 10%.

Figure 8 shows similar results, but for bias. The raw forecast exhibits a strong diurnal cycle for bias with a peak right after sunset while the collapse of the PBL is occurring, high values through the night, a minimum in the early morning, and then rapidly increasing values in the afternoon. All postprocessing methods drastically reduce the bias of the raw forecast, with AN being the best, having values close to zero for the first half of the forecast. The KF shows a nearly constant-in-time bias of around 0.1 m s⁻¹, whereas the analog-based methods' bias increases toward the end of the forecast, particularly after hour 22. Confirming what was found in Fig. 6, ANKF has, on average, a higher bias than KF and similar values to 7-Day.

The temporal evolution of rank correlation values for each method is shown in Fig. 9. All methods show a diurnal cycle, with ANKF and AN being the best, and AN being statistically significantly better than ANKF for all the 24 forecast hours. With this metric, AN provides average improvements of around 40%, 25%, 20%, and 5% with respect to Raw, 7-Day, KF, and ANKF, respectively.

c. Metrics evaluation as a function of space

An analysis of the spatial distribution of the method's skill is presented, where the metrics are computed with all the available pairs of observations and predictions at each station (i.e., N_p equal to the number of the pairs of observation and predictions available at a given station across all the times), and for all the 400 stations available. Figures 10, 11, and 12 show the results of CRMSE,

bias, and rank correlation, respectively. Each colored circle represents the value of the statistics computed with data from the station at that location. In this subsection, only Raw, KF, ANKF, and AN are analyzed, given that the pattern of performance shown previously for 7-Day is repeated when the metrics are computed for each station (not shown).

As shown in Fig. 10 (particularly for Raw and KF), CRMSE values tend to be higher in the western portion of the domain where in general the predictive skill of the forecast is lower given the challenge posed by modeling the atmospheric flow in areas with complex topography. The AN shows (with few exceptions) lighter colors corresponding to lower CRMSE values when compared to the other methods. The ANKF is a close second, followed by KF and Raw that show darker colors corresponding to higher CRMSE. This performance pattern prevails regardless of the different topographical and land-use characteristics of the locations considered, underlying the robustness of the new analog-based methods tested across a range of conditions. Similarly, Fig. 11 shows the spatial distribution of bias for the four methods. All postprocessing procedures successfully reduce the bias of the raw forecast. The KF performs better than ANKF and AN in the northeast region of the domain, whereas in the remaining portion of the domain AN is the best. At a few stations, as indicated by the red circles, the bias value is high for all the methods. These could be an indication of an observation of low quality, particularly when nearby stations show much lower biases. If such a problem exists at a particular location across different metrics and methods, then the sensors at that location may be providing unreliable observations and could be removed from the observations list; effectively, this would be equivalent to using the postprocessing methods as part of a quality-control procedure, which would improve not only the postprocessing step, but also the performance of data assimilation algorithms as well as the fidelity of verification results.

The differences across rank correlation values among the different methods as shown in Fig. 12 are less evident. ANKF and AN improve rank correlation values (i.e., darker colors) with respect to KF in the mountain regions over the western part of the domain. Overall AN exhibits the highest rank correlation values, followed closely by ANKF, and then by KF and Raw.

d. Skill scores of ANKF and AN versus KF as a function of dataset length in time

The effect of different dataset lengths on the effectiveness of the analogs found is investigated. Good analogs are those predictions from the past that are similar



FIG. 10. Spatial distribution of CRMSE (m s⁻¹) for the (a) raw forecast (Raw), (b) the KF, (c) KF run through an ordered set of analog forecasts (ANKF), and (d) the method based only on analogs (AN).

enough to the forecast that produce skillful ANKF and AN.

For this analysis, an independent NWP model dataset is used, that includes 1-yr-long (2005) time series of hourly observations and predictions from 22 stations over New Mexico. Wind forecasts with 1–24-h forecasts (at 1-h increments) were generated with the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model (MM5), run with 3.3-km horizontal grid increments and issued at 1200 UTC on a domain centered over White Sands Missile Range in New Mexico. The variables used to search for analogs include wind speed and direction, and temperature (the only variables available with this dataset). For more information on this dataset, see section 2a and Fig. 1 in Hacker and Rife (2007).

Figure 13 shows the skill score (i.e., the relative improvement in percent for a given metric) of ANKF (gray lines) and AN (black lines) with respect to KF, for both CRMSE (solid white squares) and rank correlation (solid white circles), and as a function of the dataset length (starting with the first two months and up to all 12 months of the year 2005). The skill score for bias is not shown, given that all postprocessing procedures successfully reduce the bias of the raw forecast (see Figs. 6, 8, and 11).

ANKF shows increasing improvements as the available dataset gets longer (except at 4 months for rank correlation), with maximum improvements of 11% (CRMSE) and above 8% (rank correlation) with a 12-month-long dataset. The AN shows steeper improvements than ANKF when compared to KF. Again, these improvements steadily increase with the increase of the dataset length (with the same exception at 4 months for rank correlation). With a 2-month-long dataset, the CRMSE value for AN is higher than the value for KF by more



FIG. 11. As in Fig. 10, but for bias (m s⁻¹). Red circles indicate stations for which bias is high regardless of the postprocessing method applied to the raw forecast (see discussion in text).

than 8% and by 6% for rank correlation, while with a 12-month dataset these improvements increase to 15% and more than 11%, respectively. The temporary drop of skill score at month four (April 2005) for both ANFK and AN may be due by the seasonal weather changes from winter to spring. Indeed, it may be difficult to find good analogs for forecasts issued in early spring given that the previous months belong to a different season. For the same reason, the skill score tends to flatten out or even decrease after month 10 (October 2005), because of the seasonal transition from summer to fall.

If the current forecast is for a rare prediction in the tails of the model-forecast climatological distribution, in order to find good analogs there will likely be the need to look far back in the past. On the other hand, good analogs of a forecast that is closer to climatology would likely be found with a shorter dataset, possibly with data from the same season. When longer datasets are available,

the overall chance of finding analogs similar to the forecast that would improve the performance of the analog-based methods is higher. This explains why both ANKF and AN improve their performance with respect to KF with longer datasets. The reason why improvements of AN are steeper than the ones of ANKF may resides in the differences between the two methods listed at the end of section 2f. Both methods are based on a weighted contribution of the analogs, but the way these weights are computed is different. In AN these weights are directly proportional to the closeness of an analog to a forecast. This may result in a more efficient and direct strategy to benefit from a higher quality of the selected analogs, than running KF through an ordered set of analogs and compute the weights with the Kalman gain [Eq. (4)] as in ANKF.

This finding resonates with the notion that reforecasts can greatly improve weather predictions (Hamill et al.



FIG. 12. As in Fig. 10, but for rank correlation.

2006). It is argued here that, if multiple-year datasets were available, even larger improvements than what are found in this study may be obtained with the analogbased methods. These improvements should benefit both the forecasts that are close to the climatological behavior of the model, and the rare predictions. The former would benefit from a collection of several data points belonging to the same season across different years, whereas for the latter the possibility of finding similar predictions to the forecast to be corrected would be higher if multiple years were available.

6. Summary and conclusions

Two new postprocessing methods to improve numerical weather predictions have been introduced. These methods overcome a difficulty of a postprocessing algorithm inspired by the Kalman filter (KF) and a 7-day runningmean correction (7-Day) in dealing with sudden changes of the forecast error that could be caused by rapid transitions from one weather regime to another.

The new methods rely on the analog concept. The analog of a forecast for a given location and time is defined as a past prediction that matches selected features of the current forecast. Analogs are searched across multiple physical variables and over a time window for a given location and forecast time. Good analogs are forecasts that predicted similar values and temporal trends of the forecasted quantity for which the error needs to be estimated (wind speed in this study), and for other variables that exhibit correlation to the latter (e.g., wind speed itself, wind direction, pressure, temperature, humidity, etc.). If forecasts in the past can be found that are similar to the current prediction, then the current prediction error can be inferred by analyzing the errors of the analogs, for which verifying observations are available. The two new methods are the following:



FIG. 13. Skill score (%) as a function of dataset length (month) of the KF run through an ordered set of analog forecasts (ANKF, gray lines) and the method based only on analogs (AN, black lines) against KF, and computed for CRMSE (m s⁻¹, solid white squares) and rank correlation (solid white circles).

- ANKF that consists in running KF through an ordered set of analog forecasts rather than in time. The analogs to the current forecast are ranked from the farthest (worst analog) to the closest (best analog). By running KF through the ordered (from worst to best) analogs, the correction for the current forecast gives more weight to the analog forecasts closer to it, resulting in a better correction than KF, particularly when a large variation in the day-to-day forecast error occurs.
- AN that consists of a simple method based purely on a weighted average of the observations that verified when the 10 best analog forecasts were valid [Eq. (7)]. The weights are proportional to the inverse of the distance of the analog from the forecast [as estimated with Eq. (6)], normalized with the sum of the inverse of this distance computed for each analog, and sum to 1 [Eq. (8)].

Important features of the new methods include the following:

- Analogs are searched in forecast space only (i.e., no observations are used to select the best analogs), and across different meteorological variables that exhibit correlations to the quantity of interest.
- Analogs are searched independently for each forecast time and location; the assumption is that when the space and period of the day from which analogs are selected is reduced (in other words, by "localizing" the search procedure both in space and time), also the degrees of freedom of finding matching forecasts are reduced, resulting in a more tractable problem. Moreover,

the success of the time–space localization adopted here while searching for analogs is an indication that common locations and times are good analog predictors.

• Given their design, both analog-based methods do not suffer from missing predictions or observations values, this being a clear advantage over KF and 7-Day.

The postprocessing methods have been applied to 10-m wind speed forecasts with 1–24-h forecasts (at 1-h increments) issued at 0100 UTC from the Weather Research and Forecasting modeling system run with a 3.3-km horizontal grid increment over the central United States. The methods were tested with observations from 400 surface stations for a 6-month period. The results can be summarized as follows:

- ANKF and AN are able to produce skillful corrections of the raw forecasts (Raw), even with large day-to-day changes in forecast error.
- Since AN provides a prediction based on past observations, it results in an efficient downscaling procedure that eliminates representativeness issues associated with the fact that observations and model output can be associated with quantities averaged over different spatiotemporal scales.
- All methods are successful in reducing the bias of Raw, where AN produced the lowest values (approximately 0 m s⁻¹), followed by KF, ANKF, and the 7-Day running-mean correction.
- The postprocessing methods not only reduce systematic errors but also random errors (i.e., they improve the predictive skill of the original forecast). When looking at CRMSE, ANKF and AN show improvements across all the forecast hours. Compared with KF, the improvement from ANKF and AN was in the range of 10% and 20%, respectively. There were even larger improvements with respect to Raw and 7-Day, with values close to 25% and 35%, respectively. The AN is consistently the best across the forecast period, showing statistically significant improvements over ANKF on the order of 10%. The AN is based purely on verifying observations of past predictions that are similar to the forecast (i.e., the analogs), that provide physically based insight about the atmospheric state, thus improving the predictive skill of Raw.
- For rank correlation scores, AN provides average improvements of around 40%, 25%, 20%, and 5% with respect to Raw, 7-Day, KF, and ANKF, respectively.
- Based on the Taylor diagram, AN shows a better pattern of correspondence between predictions and observations, followed by ANKF, KF, 7-Day, and Raw.
- The above relative performances prevail regardless of the time of the day, and for all of the different topographical and land-use characteristics of the

locations considered. This demonstrates the robustness of the new analog-based methods.

- When longer datasets are available, the overall performance of the analog-based methods is improved because of the increased likelihood to find good analogs in extended training periods. When extended training periods are not available, the training dataset can be extended by looking for analogs in surrounding locations, as found useful by Van den Dool (1989, 1994) and Hamill and Whitaker (2006).
- Although future studies will be needed to address the general applicability of the new analog-based methods, we believe that both ANKF and AN have the potential to be applied with success to different prediction systems and variables, given that their design does not depend on the specific variables or models used in this study. Both methods have been tested successfully with data from two independent modeling systems, as explained in sections 3a and 5d. Moreover, the KF method used in our study has been applied with success for temperature and precipitation (Roeger et al. 2003; McCollor and Stull 2008), surface winds (Roeger et al. 2003), ozone concentration (Delle Monache et al. 2006, 2008), particulate matter (Djalalova et al. 2010; Kang et al. 2010), and solar radiation (Rincon et al. 2010), and therefore we think that ANKF too has the potential to be applied successfully for other variables.

Acknowledgments. This research was supported by Xcel Energy, and by the U.S. Army Test and Evaluation Command through an Interagency Agreement with the National Science Foundation. The authors are thankful to Josh Hacker (NPS) and Daran Rife (NCAR) for providing the dataset used in the analysis presented in section 5d, and Craig Bishop (NRL) for suggesting a weighted approach for the AN method. The paper has been greatly improved by the comments and suggestions of Teddy Allen (UM); Josh Hacker, Tony Eckel, and Tom Hamill (NOAA); Thomas Warner (NCAR); and throughout and insightful revisions provided by Evan Kuchera (USAF AFWA) and an anonymous reviewer.

REFERENCES

- Bakhshaii, A., and R. Stull, 2009: Deterministic ensemble forecasts using gene-expression programming. *Wea. Forecasting*, 24, 1431–1451.
- Bergen, R. E., and R. P. Harnack, 1982: Long-range temperature prediction using a simple analog approach. *Mon. Wea. Rev.*, **110**, 1083–1099.
- Bozic, S. M., 1994: Digital and Kalman Filtering. 2nd ed. John Wiley & Sons, 160 pp.
- Carter, R. G., and R. E. Keislar, 2000: Emergency response transport forecasting using historical wind field pattern matching. *J. Appl. Meteor.*, **39**, 446–462.

- —, J. Dallavalle, and H. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, 4, 401–412.
- Chapman, W. L., and J. E. Walsh, 1991: Long-range prediction of regional sea ice anomalies in the Arctic. *Wea. Forecasting*, 6, 271–288.
- Danforth, C. M., and E. Kalnay, 2008: Using singular value decomposition to parameterize state-dependent model errors. *J. Atmos. Sci.*, 65, 1467–1478.
- —, —, and T. Miyoshi, 2007: Estimating and correcting global weather model error. *Mon. Wea. Rev.*, **135**, 281–299.
- Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. B. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman-filter predictor bias correction. J. Geophys. Res., **111**, D05308, doi:10.1029/ 2005JD006311.
- —, and Coauthors, 2008: A Kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone. *Tellus*, **60B**, 238–249.
- DelSole, T., and A. Y. Hou, 1999: Empirical correction of a dynamical model. Part I: Fundamental issues. *Mon. Wea. Rev.*, 127, 2533–2545.
- Djalalova, I., and Coauthors, 2010: Ensemble and bias-correction techniques for probabilistic forecast of surface O3 and PM2.5 during the TEXAQS-II experiment of 2006. *Atmos. Environ.*, 44, 455–467.
- Drosdowsky, W., 1994: Analog (nonlinear) forecasts of the Southern Oscillation index time series. *Wea. Forecasting*, 9, 78–84.
- Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.
- Esterle, G. R., 1992: Adaptive, self-learning statistical interpretation system for the central Asian region. *Ann. Geophys.*, 10, 924–929.
- Gao, L., H. Ren, J. Li, and J. Chou, 2006: Analogue correction method of errors and its application to numerical weather prediction. *Chin. Phys.*, 15, 882–889.
- Glahn, H., and D. Lowry, 1972: The use of model output statistics in objective weather forecasting. J. Appl. Meteor., 11, 1203– 1211.
- Hacker, J., and D. Rife, 2007: A practical approach to sequential estimation of systematic error on near-surface mesoscale grids. *Wea. Forecasting*, 22, 1257–1273.
- Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229.
- —, —, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, 87, 33–46.
- Hart, K. A., W. J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and Paralympic Winter Games. *Wea. Forecasting*, **19**, 200–218.
- Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using Kalman filter. *Wea. Forecasting*, 10, 689–707.
- Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallaville, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maxima and minima temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, 5, 128–138.
- Jolliffe, I. T., and D. B. Stephenson, 2003: Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley, 240 pp.

- Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. J. Basic Eng., 82, 35–45.
- Kang, D., R. Mathur, and S. T. Rao, 2010: Assessment of biasadjusted PM2.5 air quality forecasts over the continental United States during 2007. *Geosci. Model Dev.*, 3, 309–320.
- Leith, C. E., 1978: Objective methods for weather prediction. Annu. Rev. Fluid Mech., 10, 107–128.
- Livezey, R. E., and A. G. Barnston, 1988: Operational multifield analog/antianalog prediction system for United States seasonal temperatures. Part I: System design and winter experiments. J. Geophys. Res., 93, 10 953–10 974.
- McCollor, D., and R. Stull, 2008: Hydrometeorological accuracy enhancement via post-processing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, 23, 131–144.
- Murphy, A. H., 1988: Skill score based on the mean square error and their relationship to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.
- Radinovic, D., 1975: An analogue method for weather forecasting using the 500/1000 mb relative topography. *Mon. Wea. Rev.*, 103, 639–649.
- Ren, H., and J. Chou, 2006: Analogue correction method of errors by combining statistical and dynamical methods. *Acta Meteor. Sin.*, **20**, 367–373.
- —, and —, 2007: Strategy and methodology of dynamical analogue prediction. Sci. China Ser. D: Earth Sci., 50, 1589–1599.
- —, —, J. Huang, and P. Zhang, 2009: Theoretical basis and application of analogue-dynamical model in the Lorenz system. Adv. Atmos. Sci., 26, 67–77.
- Rincon, A., O. Jorba, and J. M. Baldasano, 2010: Development of a short-term irradiance prediction system using post-processing tools on WRF-ARW meteorological forecasts in Spain. *Extended Abstracts, European Conf. on Applied Meteorology*, Vol. 7, Zurich, Switzerland, European Meteorological Society, EMS2010-406.

- Roeger, C., R. B. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecast in mountainous terrain for application to avalanche prediction. *Wea. Forecasting*, **18**, 1140–1160.
- Skamarock, W., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/ TN-475+STR, 113 pp.
- Stensrud, D. J., and J. A. Skindlov, 1996: Gridpoint predictions of high temperature from a mesoscale model. *Wea. Forecasting*, **11**, 103–110.
- —, and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.
- Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res., 106, 7183– 7192.
- Toth, Z., 1989: Long-range weather forecasting using an analog approach. J. Climate, 2, 594–607.
- van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117**, 2230–2247.
- —, 1994: Searching for analogues, how long must we wait? *Tellus*, 46, 314–324.
- —, 2007: Empirical Methods in Short-Term Climate Prediction. Oxford University Press, 240 pp.
- Wilczak, J. M., S. A. McKeen, I. Djalalova, and G. Grell, 2006: Bias-corrected ensemble predictions of surface O3. J. Geophys. Res., 111, D23S28, doi:10.1029/2006JD007598.
- Wilks, D. S., 2006: Statistical Methods in the Atmospheric Sciences. 2nd ed. Academic Press, 627 pp.
- —, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390.
- Xavier, P. K., and B. N. Goswami, 2007: An analog method for real-time forecasting of summer monsoon subseasonal variability. *Mon. Wea. Rev.*, **135**, 4149–4160.